DOCUMENT RESUME

ED 046 460 LI 002 496

AUTHOR King, Donald W.; Neel, Peggy W.
TITLE Cost Effectiveness of On-Line Retrieval System.
INSTITUTION Westat Research, Inc., Rockville, Md.
PUB DATE Feb 71
NOTE 13p.; Paper presented at the Sixth Middle Atlantic Regional Meeting [American Chemical Society], February 3-5, 1971, Baltimore, Maryland

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS *Cost Effectiveness, Costs, *Information Retrieval, *Information Systems, Man Machine Systems, *Models, *Relevance (Information Retrieval)
IDENTIFIERS American Psychological Association

ABSTRACT

A recently developed cost-effectiveness model for on-line retrieval systems is discussed through use of an example utilizing performance results collected from several independent sources and cost data derived for a recently completed study for the American Psychological Association. One of the primary attributes of the model rests in its great flexibility in that various combinations of alternative systems and subsystems are open to comparison. Some of the systems which have been addressed include batch processing, on-line abstract and the subsystems include various levels of recall, several types of screening, and different user-system interfaces. The example chosen for discussion in this paper presents a cost-effectiveness comparison of on-line index and on-line abstract systems for various levels of demand and recall. (Author)

ED 046460

U.S. DEPARTMENT OF HEALTH, EDUCATION
& WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

COST EFFECTIVENESS OF ON-LINE RETRIEVAL SYSTEM

by

Donald W. King and Peggy W. Neel

Westat Research, Inc
Rockville, Maryland

LI 002496

Paper presented at the Sixth Middle Atlantic Regional Meeting [American Chemical Society], February 3-5, 1971, Baltimore, Maryland

COST-EFFECTIVENESS OF ON-LINE RETRIEVAL SYSTEM

An on-line system is by nature an interactive system characterized in part by the following features:

1. Speed of response
2. Ability to respond to requests for system parameters (e.g., number of documents with a given indexing)
3. Ability of natural language processing
4. Alternative of use of an intermediary. The iterative nature of an on-line system allow the direction of a particular search to change at any stage during the entire search process.

Although there has been much discussion and indecision as to appropriate measures of system effectiveness the two which follow would appear adequate for most circumstances:

1. Proportion of relevant documents retrieved
2. Total number of documents retrieved.

It has been found that a stochastic model lends itself ideally to the evaluation of retrospective search systems. In the case of an on-line system the principal components of the model are:

1. Intermediary relevance judgment, if an intermediary is used to conduct the search.
2. Query/system relevance response, which is the system's response to a series of queries.
3. Screened system relevance response, which corresponds to an intermediary's judgment (if an intermediary is used) and to the use of a document representation such as an abstract.

These components represent the various alternatives which may be combined to simulate any particular on-line retrieval system. The various sources of error may be expressed as conditional probabilities. The following notation will be used consistently:

V_r	relevant with respect to the verbalized request
$V_{\bar{r}}$	nonrelevant with respect to the verbalized request
C_r	relevant with respect to the coder's (intermediary's) interpretation
$C_{\bar{r}}$	nonrelevant with respect to the coder's (intermediary's) interpretation
R_r	relevant with respect to the system's response
$R_{\bar{r}}$	nonrelevant with respect to the system's response
S_r	relevant with respect to the screener's judgment
$S_{\bar{r}}$	nonrelevant with respect to the screener's judgment

Conditional probabilities will be designated by the standard notation P(A/B) which is read "the probability of A, given B." Thus $P(C_r/V_{\bar{r}})$ means "the probability that a document is relevant to the coder's interpretation given that it is nonrelevant with respect to the verbalized request.

The conditional probabilities used in retrospective search models are:

relevance with respect to the verbalized request	relevance with respect to coder's interpretation: $C_{\bar{r}}$	C_r
V_r	$P(C_{\bar{r}}/V_r)$	$P(C_r/V_r)$
$V_{\bar{r}}$	$P(C_{\bar{r}}/V_{\bar{r}})$	$P(C_r/V_{\bar{r}})$

relevance with respect to coder's interpretation	relevance with respect to response by system: $R_{\bar{r}}$	R_r
C_r	$P(R_{\bar{r}}/C_r)$	$P(R_r/C_r)$
$C_{\bar{r}}$	$P(R_{\bar{r}}/C_{\bar{r}})$	$P(R_r/C_{\bar{r}})$

		relevance with respect to screener's interpetation	
		S_r^-	S_r
relevance with respect to the verbalized request	V_r	$P(S_r^-/V_r)$	$P(S_r/V_r)$
	V_r^-	$P(S_r^-/V_r^-)$	$P(S_r/V_r^-)$

The conditional probailities are determined through controlled observation although in practice one is always working with relative frequencies.

So contructed, the model has the following features:

(1) It shows the following summary figures:

(a) the probability that a relevant document will be retrieved

(b) the probability that a nonrelevant document will be retrieved

(2) It shows the activities that are the principal sources of error through the entries for the conditional probabilities. Ideally, of course, all entries would be zeroes and ones, with the ones in the lower left-hand and upper right-hand corners. The amount of departure from this ideal indicates the extent of departure from perfection.

(3) The effect of error-prone components on the total output of the system can be obtained. For example, it is possible to show what effect coder interpretation errors have on system performance.

(4) It shows how specified improvement in any component will affect system output.

The model constitutes a simple application of the rules of probability and can be described mathematically as a finite Markov chain with absorbing states.

(1) $P(R_r/V_r) = P(R_r/C_r)P(C_r/V_r) + P(R_r/C_r^-)P(C_r^-/V_r)$

(2) $P(R_r/V_r^-) = P(R_r/C_r)P(C_r/V_r^-) + P(R_r/C_r^-)P(C_r^-/V_r^-)$

(3) $P(S_r, R_r/V_r) = P(S_r/V_r)P(R_r/V_r)$

(4) $P(S_r, R_r/V_r^-) = P(S_r/V_r^-)P(R_r/V_r^-)$

The notation $P(S_r, R_r/V_r)$ indicates the probability that the system has classified the document as relevant and that the screener has also classified it as such, given that the document is, in fact, relevant with respect to the verbalized request.

The theoretical recall ratio (proportion of relevant documents retrieved) is $P(S_r, R_r/V_r)$. If N_r is the number of documents in the file that is relevant to a verbalized request and $N_{\bar{r}}$ the number nonrelevant, then the theoretical precision ratio (proportion of relevant documents retrieved/nonrelevant documents retrieved) is:

$$\frac{N_r \cdot P(S_r, R_r/V_r)}{N_r \cdot P(S_r, R_r/V_r) + N_{\bar{r}} \cdot P(S_r, R_r/V_{\bar{r}})}$$

The number of documents retrieved may be found by $N_r \cdot P(S_r, R_r/V_r) + N_{\bar{r}} \cdot P(S_r, R_r/V_{\bar{r}})$

Thus far, the model results in measures of <u>effectiveness</u> rather than efficiency, since no costs have been introduced.

We at Westat, under a contract with the American Psychological Assoc developed the following generalized cost model for retrospective search systems. This model includes these subsystems:

(1) Search mode (on-line in this case)

(2) Screening processes

(3) Input (full text versus index terms and number of items input)

(4) User/system interface

(5) Method of presentation to the user

The total cost of any retrospective search system, and therefore any on-l retrieval system, is composed of three types of cost:

(1) Fixed costs associated with each subsystem

(2) Variable costs dependent on the number of items input to the syst

(3) Variable costs dependent on the number of searches conducted

Simply stated,

$C = C' + C'' X_1 + C''' X_2$

The fixed costs associated with each subsystem are:

C_1 staff, space rental, computer rental, and fixed computer storage charges for the specific computerized search system

C_2 rent, staff, and screening devices that may be used in screening the search output

C_3 input costs such as thesaurus development, staff, tape conversion, and update costs

C_4 staff, rent and sundry items involved in the user/system interface

C_5 charges for mailing the search output to users

The fixed cost element is then

$C' = C_1 + C_2 + C_3 + C_4 + C_5$

The variable costs that are dependent on the file size or number of items input to the system are:

C_6 the cost per item of indexing, abstracting, keyboarding, and any other input processing

and

$C_7 X_5$ the file loading costs, which are dependent not only on the file size, but also on the number of terms per item of input.

This cost component is then

$C'' = C_6 + C_7 X_5$

Another type of variable cost is dependent on the number of searches conducted per year. This is the annual demand for the retrospective search system. This is the most complicated element of the model because it is composed of three parts:

(1) Fixed costs per search. These are C_8 -- the set-up costs for mailing search output to users -- and C_9 -- cost of the user/system interface, i.e., the intermediary.

(2) Costs dependent on the number of items retrieved in a search.

These are C_{10} -- the computer cost of retrieving an item, C_{11} -- the cost of printing an item, and C_{12} -- the cost of screening each item retrieved.

(3) Costs dependent on the number of items mailed per search.

This is C_{13} -- the cost of actually mailing the search output to the user.

This cost component can be expressed as:

$$C''' = C_8 + C_9 + X_3(C_{10} + C_{11} + C_{12}) + X_4C_{13}$$

Combining the elements of the cost equation, we have:

$$C = C_1 + C_2 + C_3 + C_4 + C_5 + X_1(C_6 + C_7X_5) + X_2 [C_8 + C_9 + X_3(C_{10} + C_{11} + C_{12}) + X_4C_{13}]$$

where

X_1 = number of items input

X_2 = number of searches conducted

X_3 = number of items retrieved per search

X_4 = number of items mailed per search

X_5 = number of terms in authority list

C_1 = fixed cost associated with computing

C_2 = fixed cost associated with screening

C_3 = fixed cost associated with input

C_4 = fixed cost associated with user/system interface

C_5 = fixed cost associated with mailing results

C_6 = total input cost per item

C_7 = total file loading cost per item per term

C_8 = fixed cost of mailing per search

C_9 = fixed cost of user/system interface per search

C_{10} = computer retrieval cost per item retrieved

C_{11} = computer printing cost per item retrieved

C_{12} = screening cost per item retrieved

C_{13} = mailing cost per item mailed

C = total annual cost

This general equation can be used to estimate costs of potential search systems as well as to compare the cost/effectiveness trade-off of system alternatives.

Table 1 shows some various alternatives for on-line retrospective searching along with associated effectiveness probabilities and cost figures.

Using the figures noted in Table 1 and applying the model as outlined results in summary figures such as those shown in Table 2.

Once effectiveness figures and system costs have been determined and the summary figures in Table 2 calculated, cost/effectiveness decision-making may begin. The weight to be assigned to each factor, of course, will depend upon specific system and organizational parameters, goals, objectives and operational constraints. It is most important that these factors be clearly understood by the cost/effectiveness team before the decision-making process is undertaken.

Table 1 - Alternative On-line Retrospective Searching Processes: Effectiveness Probabilities and Costs

Alternatives	Effectiveness Probabilities	Fixed Costs	Variable Costs
Intermediary interpretation By letter	$P(C_r \mid V_r) = 0.95$ $P(C_r \mid V_{\bar{r}}) = 0.0000034$	C_4 = \$500	C_9 = \$11.25/search
Searching On-line index Level 1 Level 2 Level 3 Level 4 Index input	$P(R_r \mid C_r)$ = 0.40, 0.60, 0.80, 1.00 $P(R_r \mid C_{\bar{r}})$ = 0.000079, 0.000227, 0.000626, 0.001656	C_1 = \$68,200 C_3 = \$ 6,925	$C_{10} + C_{11}$ = \$0.4415/item retrieved = \$0.3194/item retrieved = \$0.2494/item retrieved = \$0.2218/item retrieved $C_6 + C_7 X_5$ = \$0.6755/item input
Searching On-line abstract Level 1 Level 2 Level 3 Level 4 Abstract input	$P(R_r \mid C_r)$ = 0.40, 0.60, 0.80, 1.00 $P(R_r \mid C_{\bar{r}})$ = 0.000061, 0.000132, 0.000417, 0.001280	C_1 = \$191,750 C_3 = \$ 7,350	$C_{10} + C_{11}$ = \$0.6768/item retrieved = \$0.4029/item retrieved = \$0.3756/item retrieved = \$0.3235/item retrieved $C_6 + C_7 X_5$ = \$1.6825/item input

Table 1 (Continued) - Alternative On-line Retrospective Searching Processes: Effective Probabilities and Costs

Alternatives	Effectiveness Probabilities	Fixed Costs	Variable Costs
Screening			
Mail		C_5 = \$2,500	C_8 = \$0.20/search
No screen	$P(S_r \mid V_r) = 1.00$ $P(S_r \mid V_{\bar{r}}) = 1.00$	C_2 = \$0	C_{13} = \$0.002/item sent
	= 0.555		C_{12} = \$0
Loose-on titles	= 0.875	= \$0	= \$0.04/item retrieved

X_1 = 100,000 items over a 4-year period.

X_2 = 4,000 searches per year.

X_5 = 1,000 terms on authority list.

Table 2 Summary Retrieval and Cost Figures Associated with Combinations of Alternative Searching Processes

Recall level	Alternative search system	Recall	No. rel. retr.	No. retr. (X_3)	No. sent (X_4)	Total system cost	System cost/ search	System cost/ rel. retr.
0.40	On-line index/ no screen	0.38	19.0	27.0	27.0	$188,708	$ 47.18	$2.48
	On-line abstract/ no screen	0.38	19.0	25.3	25.3	$358,642	$ 89.66	$4.72
	On-line index/ screen on titles	0.33	16.6	27.0	21.1	$192,984	$ 48.24	$2.91
	On-line abstract/ screen on titles	0.33	16.6	25.3	20.1	$362,662	$ 90.67	$5.46
0.60	On-line index/ no screen	0.57	28.5	51.4	51.4	$206,891	$ 51.72	$1.81
	On-line abstract/ no screen	0.57	28.5	41.9	41.9	$354,649	$ 88.66	$3.11
	On-line index/ screen on titles	0.50	24.9	51.4	37.6	$214,992	$ 53.75	$2.16
	On-line abstract/ screen on titles	0.50	24.9	41.9	31.5	$364,444	$ 91.11	$3.66
0.80	On-line index/ no screen	0.76	38.0	100.9	100.9	$242,277	$ 60.57	$1.59
	On-line abstract/ no screen	0. 76	38.0	79.9	79.9	$410,643	$102.66	$2.70
	On-line index/ screen on titles	0.67	33.3	100.9	68.1	$258,159	$ 64.54	$1.94
	On-line abstract/ screen on titles	0.67	33.3	79.9	56.5	$423,879	$105.97	$3.18

Recall level	Alternative search system	Recall	No. rel. retr.	No. retr. (X_3)	No. sent (X_4)	Total system cost	System cost/ search	System cost/ rel. retr.
1.00	On-line index/ no screen	0.95	47.5	213.3	213.3	$331,758	$ 82.94	$1.75
	On-line abstract/ no screen	0.95	47.5	175.4	175.4	$518,333	$129.58	$2.73
	On-line index/ screen on titles	0.83	41.6	213.3	133.5	$365,248	$ 91.31	$2.20
	On-line abstract/ screen on titles	0.83	41.6	175.4	112.6	$545,895	$136.47	$2.96

EXAMPLE

Alternative	Effectiveness Figures	
letter	$C_r/V_r = .95$	$C_r/V_{\bar{r}} = .0000034$
online abstract level 3 abstract input	$R_r/C_r = .80$	$R_r/C_{\bar{r}} = .000417$
loose screening on titles	$S_r/V_r = .875$	$S_r/V_{\bar{r}} = .555$

Cost Figures

Fixed

$C_4 = \$500$
$C_1 = \$191,750$
$C_3 = \$7,350$
$C_2 = \$0$
$C_5 = \$2,500$
$C_8 = \$ .20$ search

Variable

$C_9 = \$11.25$ search
$C_{10} + C_{11} = \$0.3756$/item retrieved
$C_6 + C_7X_5 = \$1.6825$/item input
$C_{12} = \$0.04$/item retrieved
$C_{13} = \$ .002$/item sent

Effectiveness Model

$P(R_r/V_r) = P(R_r/C_r)P(C_r/V_r) + P(R_r/C_{\bar{r}})P(C_{\bar{r}}/V_r) = (.80)(.95) + (.000417)(.05)$ or $P(R_r/V_r) = .76002085$

$P(R_r/V_{\bar{r}}) = P(R_r/C_r)P(C_r/V_{\bar{r}}) + P(R_r/C_{\bar{r}})P(C_{\bar{r}}/V_{\bar{r}}) = (.80)(.0000034) + (.000417)(.9999966)$ or $P(R_r/V_{\bar{r}}) = .00041734$

$P(S_r, R_r/V_r) = P(S_r/V_r)P(R_r/V_r) = (.875)(.76002085)$ or $P(S_r, R_r/V_r) = .66501824375$

$P(S_r, R_r/V_{\bar{r}}) = P(S_r/V_{\bar{r}})P(R_r/V_{\bar{r}}) = (.555)(.00041734)$ or $P(S_r, R_r/V_{\bar{r}}) = .0002316237$

50 relevant documents
~~499,050~~ 99,950 nonrelevant documents

recall = .67
No. of relevant documents retrieved and passed by screener = 33
No. of nonrelevant documents retrieved and passed by screener = 23
No. of documents sent to user = 56

$C = C_1 + C_2 + C_3 + C_4 + C_5 + X_1(C_6 + C_7X_5) + X_2[C_8 + C_9 + X_3(C_{10} + C_{11} + C_{12}) + X_4C_{13}]$

$= \$202,100 + 24,000 (\$1.6825) + 4,000 [\$ .20 + \$11.25 + 79.9 (\$ .3756 + \$ .04) + 56.5 (\$ .002)]$
$= \$202,100 + \$42,062 + 4,000 [\$ .20 + \$11.25 + \$33.206 + \$ .113]$
$= \$202,100 + \$244,162 + 4,000 [\$44.769]$
$= \$202,100 + \$244,162 + \$179,077.76$
$C = \$423,000$